Speaker Diarization of Multi-party Conversations using Participants Role Information: Political Debates and Professional Meetings

Fabio Valente[1] and Alessandro Vinciarelli[1,2]

[1] Idiap Research Institute, Martigny (Switzerland)
[2] University of Glasgow, Glasgow (United Kingdom)
fvalente@idiap.ch; Alessandro.Vinciarelli@glasgow.ac.uk

1 Introduction

Speaker Diarization aims at inferring *who spoke when* in an audio stream and involves two simultaneous unsupervised tasks: (1) the estimation of the number of speakers, and (2) the association of speech segments to each speaker. Most of the recent efforts in the domain have addressed the problem using machine learning techniques or statistical methods (for a review see [11]) ignoring the fact that the data consists of instances of human conversations.

When humans want to use language to communicate orally with each other, they are faced to a coordination problem. "*Avoidance of collision is one obvious ground for this coordination of actions between the participants. In order to coordinate efficiently and successfully, they will therefore have to agree to follow certain rules of interaction*" [8]. One such rule is that no one monopolizes the floor but the participants take turns to speak. This concept is called *turn-taking*. The computational linguistic literature is rich on the analysis of human conversations; the seminal work of [9] shows that conversations obey to predictable interactions pattern between participants and a speaker turn is related in predictable ways to the previous and next turn and follows a structure similar to a grammar. In between the social phenomena that regulates the turns in a conversation, lot of attention has been devoted to roles. In fact people interact in different ways depending on the context of the environment but "*Their interactions involve behaviors associated with defined statuses and particular roles. These statuses and roles help to pattern our social interactions and provide predictability*" [10].

Only recently it has been shown that the turn-taking behavior can be statistically modeled and used to automatically classify a certain number of characteristics in groups conversations like roles. Examples include the automatic recognition of roles in meetings recordings like CMU or AMIDA recordings [2, 4], the recognition of participant seniority (professor, phd or graduate student) in the ICSI meeting data set [6] and the recognition of functional roles in the MSC corpus [3, 15]. Typically those studies are based on the use of statistical

classifiers trained on a set of automatically or semi-automatically derived audio features including the speaker turn durations, the overlap between speakers and the speaker turn statistics. They assume that the participants interactions and specifically the turn-taking patterns can be statistically modeled and provide enough information for recognizing the role of each speaker in the conversation.

This work investigates whether the use of the statistical information derived from roles can reversely increase the performance of conventional audio processing systems like diarization. In details, this work discusses the use of turn-taking information induced by the roles that participants have in the discussion as prior information in the speaker diarization systems. Previous attempts have used participant interaction patterns to improve the diarization performance, e.g. [5], however this information was not induced by, or put in relation with, any social phenomena. In this work, we make the following hypothesis: 1) the turn-taking patterns are conditioned on the role that each speaker has in the conversation, 2) they can be estimated on an independent development data set.

We propose to model the speaker sequence using N-gram of speaker roles. N-gram models can be then combined with the acoustic information coming from MFCC features. The approach is largely inspired by the current Automatic Speech Recognition (ASR) framework where the acoustic information from the signal, i.e., the acoustic score, is combined with the prior knowledge from the language, i.e., the language model. The most common form of language model is represented by words N-gram. In a similar way, given a mapping speakers to roles, N-gram models can encode the statistical information on how the participants take turns in the conversation.

The investigation is carried on two very different dataset, the first one is composed of political debates recorded with close-talk high quality microphones while the second one is composed of professional meetings recorded with far-field low quality microphones. The use of those datasets aim at studying how those findings generalizes across different types of conversations and different acoustic conditions. Let us briefly describe those datasets in the following.

2 Data description

The first dataset used for this study consists of political debates [14] that represent an excellent resource for their realism. In contrast with other benchmarks, political debates are real-world data. Debate participants do not act in a simulated social context, but participate in an event that has a major impact on their real life (for example, in terms of results at the elections). Thus, even if the debate format imposes some constraints, the participants are moved by real motivations leading to highly spontaneous social behavior.

Each debate revolves around a yes/no question like “Are you favorable to new laws on education ?”. The participants state their answer (yes or no) at the beginning of the debate and do not change it during the discussion. Each debate involves a moderator and a variable number of guests (four or more). The dataset is annotated in terms the role that each participant has in the discussion, i. e.

moderator or guests. All debates include one moderator expected to ensure that all participants have at disposition the same amount of time for expressing their opinion. Furthermore, the moderator intervenes whenever the debate becomes too heated and people tend to interrupt one another or to talk together. The guests are labeled in terms of groups according to how they answer to the central question of the debate. Participants belonging to the same group agree with one another, while participants belonging to different groups disagree with one another. The dataset is divided in two non-overlapping parts, a development dataset (composed of 25 debates for a total of 17 hours and 2600 speaker turns) and a test dataset (composed of 25 debates for a total of 15 hours and 2500 speaker turns).

The second dataset is based on the AMI meeting database [7], a collection of 138 meetings recorded with distant microphones for approximatively 100 hours of speech, manually annotated at different levels (roles, speaking time, words, dialog acts). Each meeting consists of a scenario discussion in between four participants where each participant has a given role: project manager PM, user interface expert UI, marketing expert ME and industrial designer ID. The scenario consists in four employes of an electronic company that develop a new type of television remote controller. The meeting is supervised by the project manager. The dataset is divided in two non-overlapping parts, a development data set (118 meetings) and a test set (20 meetings).

3 Turn-Taking patterns and Roles

Let us formalize the turn-taking and role informations as follows. For each recording the following triplets are available:

$$T = \{(t_1, \Delta t_1, s_1), ..., (t_N, \Delta t_N, s_N)\} \tag{1}$$

where t_n is the beginning time of the n-th turn, Δt_n is its duration, s_n is the speaker associated with the turn and N is the total number of turns in the recording. The begin of the turn corresponds to the time at which the speaker s_n grabs the floor of the discussion and the length Δt_N corresponds to the time during which s_n holds the floor.

Each participant is labeled according to the role he or she has in the recording and the mapping between each speaker and his/her role is given by the function $\varphi(S) \rightarrow R$. In case of debates the roles are moderator m, or guest g. Guests are furthermore labeled in two groups $g1$ and $g2$ according to their agreement/disagreement thus the space of roles is given by $R = \{m, g1, g2\}$. On the other hand , in case of meetings, the space of roles is given by $R = \{PM, UI, ME, ID\}$.

The sequence of speakers $S = \{s_1, ..., s_n\}$ can be statistically modeled as a first-order Markov chain in which the probability of the participant s_n speaking after the participant s_{n-1} is regulated by their respective roles $\varphi(s_n)$ and $\varphi(s_{n-1})$ (see [13]).

Table 1 represents the conditional probability $p(\varphi(s_n)|\varphi(s_{n-1}))$ of a speaker role conditioned to the role of the previous speaker on the development dataset in case of debates while Table 2 represent the same quantities in case of meeting recordings. Those statistics are obtained disregarding overlapping speech regions (including back-channels).

	Moderator	Group 1	Group 2
Moderator	0	0.51	0.49
Group 1	0.68	0.06	0.26
Group 2	0.67	0.25	0.08

Table 1. Transition matrix between roles estimated on the debates development data set.

	PM	UI	ME	ID
PM	0	0.34	0.31	0.35
UI	0.39	0	0.30	0.31
ME	0.43	0.28	0	0.29
ID	0.41	0.29	0.30	0

Table 2. Transition matrix between roles estimated on the meetings development data set.

Tables 1 and 2 can be interpreted in straightforward way. In case of debates, the moderator aims at sharing the available time in between the two groups and this is reflected in the fact that $p(g1|m)$ is approximatively equal to $p(g2|m)$ as well as $p(m|g1)$ is approximatively equal to $p(m|g2)$. On the other hand speakers with different opinions are more likely to take turn (on average) after a speaker they disagree with and this explains why $p(g2|g1)$ and $p(g1|g2)$ are considerably higher then $p(g1|g1)$ and $p(g2|g2)$. The probability $p(m|m)$ is equal to zero as there is only one moderator in each debate.

In case of meetings the Program Manager acts as moderator aiming at sharing the time in between the other participants; similarly the probability that a participant will take turn after the Program Manager is higher then the probability of taking turn after a non-chairperson participants.

In other words, the possible speaker sequences $S = \{s_1, ..., s_N\}$ in a conversations are not all equally probable and their probability can be simply estimated as:

$$p(S) = p(s_1, ..., s_n) = p(\varphi(s_1), ..., \varphi(s_n)) = p(\varphi(s_0)) \prod_{i=1}^{N} p(\varphi(s_n)|\varphi(s_{n-1})) \tag{2}$$

where $p(\varphi(s_n)|\varphi(s_{n-1}))$ are elements of the matrix (1) and $p(\varphi(s_0))$ is the probability of the role associated with the speaker that opens the discussion. In

the most general case the sequence S can be modeled using an N-gram, i.e.:

$$p(S) = p(s_1, ..., s_n) = p(\varphi(s_1), ..., \varphi(s_n)) =$$
$$= p(\varphi(s_1), ..., \varphi(s_p)) \prod_{n=p}^{N} p(\varphi(s_n)|\varphi(s_{n-1}), ..., \varphi(s_{n-p})) \quad (3)$$

where the probability of a speaker taking the n-th turn is conditioned to the role of the previous p speakers taking turns before him. Those N-gram models will be referred as *speaker role N-gram* and the paper will investigate how this information can be included as prior knowledge in a speaker diarization system.

4 Speaker diarization system

Speaker Diarization is the task that aims at inferring *who spoke when* in an audio stream. The system used here is a state-of-the-art system described in [12] and briefly summarized in the following.

Acoustic features consist of 19 MFCC coefficients extracted using a 30ms window shifted by 10ms. After speech/non-speech segmentation and rejection of non-speech regions, the acoustic features $X = \{x_1, \ldots, x_T\}$ are uniformly segmented into chunks of 250ms. Then hierarchical agglomerative clustering is performed grouping together speech segments according to a distance inspired from information theory and the clustering stops when a criterion based on Normalized Mutual Information (NMI) is met (see [12] for details). This produces an estimate of the number of participants in the debate and a partition of the data in clusters, i. e., it associates each acoustic vector x_t to a speaker s. As the diarization system classifies silence regions as non-speech, the actual turn-taking can be obtained bridging together consecutive speech segments from the same speaker separated by silence regions. For instance, the turns can simply be obtained bridging the silence regions that separates the three utterances spoken by the first speaker.

We refer this initial segmentation into speakers as T^*:

$$T^* = \{(t_1^*, \Delta t_1^*, s_1^*), ..., (t_N^*, \Delta t_N^*, s_N^*)\} \quad (4)$$

After clustering, the speaker sequence is re-estimated using an ergodic Hidden Markov Model/Gaussian Mixture Model where each state represents a speaker. The emission probabilities are modeled as GMMs trained using acoustic vectors x_t assigned to speaker s. Each state enforces a minimum duration constraint. This step aims at refining the data partition obtained by the agglomerative clustering and improving the speaker segment boundaries [11].

The decoding is performed using a conventional Viterbi algorithm, i. e. the optimal speaker sequence $\mathbf{S}^* = (s_1, s_2, ..., s_N)$ is obtained maximizing the following likelihood:

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} \log\, p(X|S) \quad (5)$$

The emission probability $p(x_t|s_t)$ of the acoustic vector x_t conditioned to speakers s_t is:

$$\log\, p(x_t|s_t) = \log \sum_r w^r_{s_t} \mathcal{N}(x_t, \mu^r_{s_t}, \Sigma^r_{s_t})$$

where $\mathcal{N}(.)$ is the Gaussian pdf; $w^r_{s_t}, \mu^r_{s_t}, \Sigma^r_{s_t}$ are weights, means and covariance matrix corresponding to speaker model s_t. The output of the decoding step is a sequence of speakers with their associated speaking time.

Let us report the performance of this system on the meetings and the debates that compose the test data set. The most common metric for assessing diarization performances is the Diarization Error Rate [3] which is composed by speech/non-speech and speaker errors. As the same speech/non-speech segmentation is used across experiments, in the following only the speaker error is reported. Table 3 reports the speaker error in case of a-priori known number of speakers K. It can

	Debates	Meetings
Speaker Error	6.2%	14.4%

Table 3. Speaker Error reported on the test data set in case of debates and meetings.

be notice from table 3 that the diarization performance is significantly worst in case of meetings because the audio is recorded with far field microphones while in case of debates the audio is acquired using close talk microphones.

5 Speaker-turns based diarization

The decoding step 5 only depends on the acoustic score $p(X|S)$ (see Eq. (5)) and completely neglects the fact that not all speaker sequences S have the same probability. In section 3, we discussed that the roles regulate the way speakers take turns and the probability of a given speaker sequence can be estimated using Eq. (3). It is thus straightforward to extend the objective function (see Eq. 5) in order to include this type of information i. e.:

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} \log\, p(X|S)p(S) = \arg\max_{\mathbf{S}} \log\, p(X|S)p(\varphi(S)) \qquad (6)$$

In other words, the optimal speaker sequence (and the associated speaker times) can be obtained combining the evidence from the acoustic score $p(X|S)$ together with the prior probability of a given sequence $p(S)$. This is somehow similar to what is done in Automatic Speech Recognition (ASR) where sentences (i. e. word sequences) are recognized combining acoustic information together with linguistic information captured in the language model. Looking at Eq. (6), it is possible to notice that while the acoustic score $p(X|S)$ is modeled using a

[3] `http://www.itl.nist.gov/iad/mig/tests/rt/`

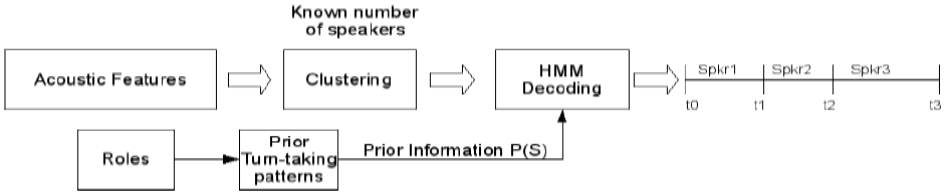

Fig. 1. Schematic representation of the proposed system in case scenario 1 (known number of speakers and roles): the clustering stops when the known number of clusters is obtained; Speaker decoding is done combining the acoustic information with prior turn-taking information induced by participants role.

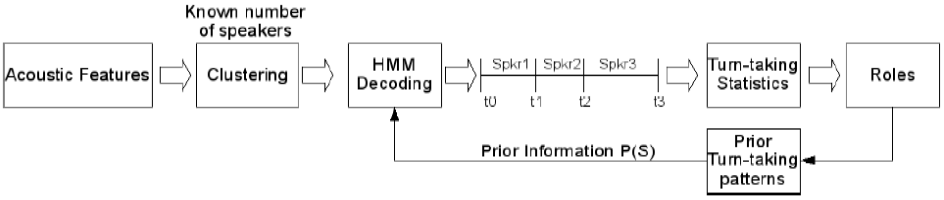

Fig. 2. Schematic representation of the proposed system in case scenario 2 (known number of speakers and unknown roles): the clustering stops when the known number of clusters is obtained; turn-taking statistics obtained from the diarization output are used to recognize speaker roles. Roles are then used to compute the prior probability of a speaker sequences $P(S)$ which is used then in the diarization system.

probability density function, i. e. a GMM, $p(S)$ is a probability; as in ASR, we introduce a factor λ tuned on the development data set to scale $P(S)$ at the same order of magnitude of $p(X|S)$ and an insertion penalty:

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} [\log p(X|S)\, p(\varphi(S))^{\lambda}] \tag{7}$$

Eq (7) can be solved using a Viterbi decoder that includes the prior probability of different speaker sequences. The development data set is used to estimate the probabilities $p(\varphi(S)$ and the scaling factor λ as well as the decoder insertion penalty. Performances are reported on the evaluation data set. In the most general case, the speaker roles are unknown. To incrementally study the integration of prior information $p(S)$, two different case scenarios are proposed.

5.1 Case 1

The number of participants K (thus speakers) in the debate is known as well as the mapping speakers-role $\varphi(.)$. The entire process is schematically depicted in Figure 1.

Those assumptions significantly simplify the problem. The clustering stops whenever the number of clusters is equal to the actual number of participants

in the recording and the mapping speaker-role is obtained from the manual reference thus the prior $P(S)$ can be directly estimated from Eq. (3). Table 4 reports the speaker error obtained with conventional decoding and with role-based decoding. The inclusion of the prior information reduces the speaker error from 6.2% to 4.6% i. e. a relative improvement of 25% for debates recordings and from 14.4% to 11.5% for meeting recordings, i.e., a 19% relative improvement. The improvements are verified on all the recordings from the data set. The largest

Table 4. Speaker Error obtained using unigrams, bigrams and trigrams in case scenario 1. In brackets the relative improvement is reported w. r. t. the baseline where no prior information is available.

Prior	$P(\varphi(s_n)$	$P(\varphi(s_n)\|\varphi(s_{n-1}))$	$P(\varphi(s_n)\|\varphi(s_{n-1},\varphi(s_{n-2}))$
Debates - Sp. Err.	5.8 (+6%)	4.6 (+25%)	4.6 (+25%)
Meetings - Sp. Err.	13.8% (+4%)	11.8% (+18%)	11.5% (+19%)

reduction in the error rate is obtained using a bigram model, i.e., conditioning the turn to the role of the previous speaker. The use of trigram models only marginally improve over the bigram. It is interesting to notice that the approach appears effective on different type of acoustic conditions (far-field and close talk audio) and on different type of data, political debates and professional meetings. This suggest that the method could be applied to any type of multi-party conversation once a mapping from speakers to role is known.

5.2 Case 2

In this case we assume that the number of participants K in the debate is known but the mapping speakers-role $\varphi^*(.)$ is estimated from the segmentation T^*. The entire process is schematically depicted in Figure 2. As before, the clustering stops whenever the number of clusters is equal to the actual number of participants in the recording producing an initial solution T^*. The mapping speakers-role $\varphi^*()$ is estimated from the segmentation T^* using the following maximization:

$$\varphi^* = \arg\max_{\varphi} p(\varphi(s_0^*)) \prod_{n=1}^{N} p(\varphi(s_n^*)|\varphi(s_{n-1}^*)). \tag{8}$$

The optimization (8) is performed exhaustively searching the space of possible mappings speakers-roles, i. e., $\varphi(\{s_k\}) \rightarrow \{R\}$ and selecting the mapping that maximize the probability of the speaker sequence s^*, i. e., Eq. (8). Table 5 reports the speaker error obtained with conventional decoding and with role-based decoding. The inclusion of the prior information reduces the speaker error from 6.2% to 4.9% i. e. a relative improvement of 20% for debates recordings and from 14.4% to 11.9% for meeting recordings, i.e., a 17% relative improvement. Again the largest reduction in the error rate is obtained using a bigram

Table 5. Speaker Error obtained unigrams, bigrams and trigrams in case scenario 2. In brackets the relative improvement is reported w. r. t. the baseline where no prior information is available.

Prior	$P(\varphi(s_n))$	$P(\varphi(s_n)\|\varphi(s_{n-1}))$	$P(\varphi(s_n)\|\varphi(s_{n-1},\varphi(s_{n-2}))$
Debates - Sp. Err.	5.9 (+6%)	4.9 (+20%)	4.9 (+20%)
Meetings - Sp. Err.	14.4% (+3%)	12.0% (+16%)	11.9% (+17%)

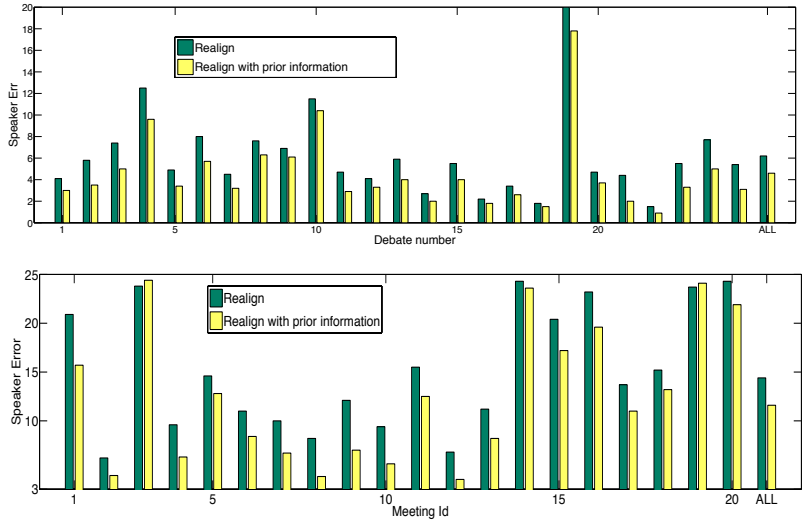

Fig. 3. Speaker error obtained using realignment with and without prior information for the 25 recordings that compose the debates test data set (top figure) and for the 20 recordings that compose the meeting test data set. The speaker error is reduced on all the debates as well as on 18 meetings out of 20.

model, i.e., conditioning the turn to the role of the previous speaker. The use of trigram models only marginally improve over the bigram. Improvements are slightly smaller compared to those obtained in Case 1 because of errors that occurs when roles are estimated using Eq. 5.

Figure 3 plots the speaker error with and without prior information for the 25 recordings that compose the test data set in Case 2. The proposed approach reduces the speaker error on 23 out of 25 debates in Case 2. The error does not decrease in two recordings with high speaker error. In Case 1 and Case 2 (not plotted), the improvements are verified on all the 25 recordings. We do not verify a degradation in performance in any recording.

Let us now investigate the differences between the systems outputs. Figures 4 plots the relative amount of total speaker time correctly attributed to each of the four roles by the baseline diarization and the proposed technique. Those statistics are averaged over the entire test set and normalized dividing by the

total speaker time. The largest improvement in performance comes from the time correctly attributed to the speakers labeled as PM in meetings (see figure 4 (a)) and as moderator in debates (see figure 4 (b)). In the psychology literature, those roles (moderator and project manager) can be associated with the *gatekeeper* (see [1]), i.e., the speaker that encorages and regulates the discussion. In other words, most of the improvements comes from the speech attributed to the *gatekeeper* of the discussion rather then from speech attributed to the other roles.

Further analysis shows that the proposed method outperforms the baseline especially on short turns where the acoustic score may not provide enough information to assign the segment to a given speaker.

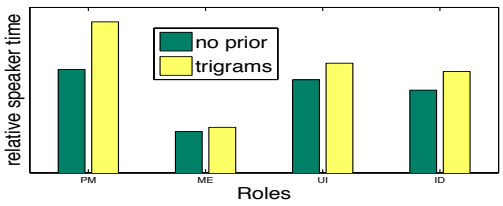

Fig. 4. Relative amount of speaker time correctly attributed to each of the four speakers labeled according to their roles by the baseline diarization and the proposed technique in case 2 in case of meeting recordings. Statistics are averaged over the entire test set.

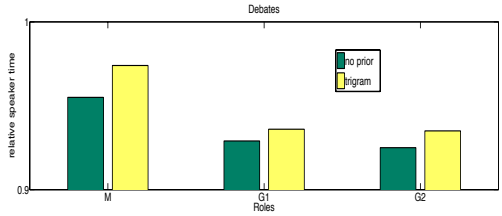

Fig. 5. Relative amount of speaker time correctly attributed to each of the four speakers labeled according to their roles by the baseline diarization and the proposed technique in case 2 in case of meeting recordings. Statistics are averaged over the entire test set.

6 Discussions

A large body of recent works has focused on the recognition of roles in multi-party discussions. Turn-taking patterns, i.e. the tendency of participants to interact or to react to certain persons rather then others, represents a powerful cue for inferring the role that each speaker has in a discussion [3, 15]. Speaker diarization represents a key techonolgy for automatic turns extraction.

This work discusses the use of turn-taking patterns as a priori information in diarization systems. In contrary to related works [5], the patterns are explicitly put in relation with the roles that each speaker has in the discussions and they are estimated on an independent development data set. Experiments are carried out on political debates and professional meeting recordings. Those two datasets have different acoustic conditions (close talk speech for the first and far-field speech for the second) and represent different type of conversations (competitive debate in the first case versus professional collaborative meeting in the latter).

Results show that whenever the number of participants in the discussion as well as their roles are known the speaker error is reduced by 25% in case of debates and by 20% in case of meetings; whenever the second one is not available the improvements are 20% in case of debates and 17% in case of meetings. In summary the proposed method seem to reduce consistently the speaker error across different types of conversations and different acoustic conditions. The largest error reduction is obtained when bigram of roles are used; the use of trigrams marginally reduces the total error respect to the bigrams.

The largest part of the improvements come from speech attributed to the debate moderator or the meeting program manager; those roles can be associated with the *gatekeeper* (according to the social role coding scheme [1]), i.e., the speaker that encorages and regulates the discussion.

References

1. R.F. Bales. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley, 1950.
2. S. Banerjee and A.I. Rudnicky. Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of International Conference on Spoken Language Processing*, number 2-3, pages 221–231, 2004.
3. W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 271–278, 2007.
4. N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli. Role recognition for meeting participants: an approach based on lexical information and Social Network Analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696, 2008.
5. K.J. Han and S.S. Narayanan. Improved speaker diarization of meeting speech with recurrent selection of representative speech segments and participant interaction pattern modeling. In *Proceedings of Interspeech*, pages 1067–1070, 2009.
6. K. Laskowski, M. Ostendorf, and T. Schultz. Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In *In proceedings of the 9th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, pages 148–155, June 2008.
7. I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos. The AMI meeting corpus. In *Proceedings of the International Conference on Methods and Techniques in Behavioral Research*, volume 88, 2005.

8. B. Oreström. *Turn-taking in English conversation.* Krieger Pub Co, 1983.
9. H. Sacks, E.A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735, 1974.
10. H. Tischler. *Introduction to sociology.* Harcourt Barce College Publishers, 1990.
11. S.E.E Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565, 2006.
12. D. Vijayasenan, F. Valente, and H. Bourlard. An information theoretic approach to speaker diarization of meeting data. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1382–1393, 2009.
13. A. Vinciarelli. Capturing order in social interactions. *IEEE Signal Processing Magazine*, 26(5):133–152, 2009.
14. A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *Proceedings of International Workshop on Social Signal Processing*, pages 1–4, 2009.
15. M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proceedings of International Conference on Mutlimodal Interfaces*, pages 47–54, 2006.